

PERBANDINGAN ALGORITMA C4.5 DAN NEURAL NETWORK UNTUK MEMPREDIKSI HASIL PEMILU LEGISLATIF DKI JAKARTA

Mohammad Badrul

Program Studi Sistem Informasi
STMIK Nusa Mandiri Jakarta

Jl. Damai No. 8 Warung Jati Barat (Margasatwa) Jakarta Selatan. Telp. (021) 78839513
mohammad.mbl@nusamandiri.ac.id

ABSTRACT

General elections are a means of implementation of the sovereignty of the people in the Unitary State of Indonesia based on Pancasila and 1945 Constitution. Elections held in Indonesia is to choose the leadership of both the president and vice president, member of parliament, parliament, and the DPD. In this study comparison of data mining methods, namely C4.5 and neural network algorithm is applied to both the data legislative candidates to be elected to the legislature and were not selected. C4.5 algorithm is one of the algorithms in a decision tree method that converts the data into a decision tree using entropy calculation formula. While the neural network algorithm is a method like human neurons to find the best path. From the test results to measure the performance of both methods using cross validation test method, confusion matrix and ROC curves is known that the neural network has the highest accuracy value which is equal to 98.50%, followed by the C4.5 algorithm method with 97.84% accuracy values. AUC values for the neural network method showed the highest value of 0.982 and a decision tree algorithm with a value of 0.970.

Kata Kunci: Elections, C4.5 algorithm, neural network algorithm, accuracy.

PENDAHULUAN

Pemilihan umum (Pemilu) adalah sarana pelaksanaan kedaulatan rakyat dalam negara kesatuan Republik Indonesia yang berdasarkan Pancasila dan Undang-Undang Dasar 1945 (Undang-Undang RI No.10 Tahun 2008 tentang Pemilihan umum). Pemilihan umum adalah salah satu pilar utama untuk memilih pemimpin dari sebuah demokrasi atau bisa disebut yang terutama (Santoso, 2004). Pemilu merupakan sarana yang sangat penting bagi terselenggaranya sistem politik yang demokratis. Karena itu, tidak mengherankan banyak negara yang ingin disebut sebagai negara demokratis menggunakan pemilu sebagai mekanisme membangun legitimasi. Pemilu bertujuan untuk memilih anggota Dewan Perwakilan Rakyat (DPR), Dewan Perwakilan Rakyat Daerah (DPRD) provinsi, dan Dewan Perwakilan Rakyat Daerah (DPRD) kabupaten/kota yang dilaksanakan dengan sistem proporsional terbuka (Undang-Undang RI No.10 Tahun 2008 tentang Pemilihan umum). Dengan sistem pemilu langsung dan jumlah partai yang besar maka pemilu legislatif memberikan peluang yang besar pula bagi rakyat Indonesia untuk berkompetisi menaikkan diri menjadi anggota

legislatif. Pemilu legislatif tahun 2009 diikuti sebanyak 44 partai yang terdiri dari partai nasional dan partai lokal. Pemilu Legislatif DKI Jakarta Tahun 2009 terdapat 2.268 calon anggota Dewan Perwakilan Rakyat Daerah (DPRD) dari 44 partai yang akan bersaing memperebutkan 94 kursi anggota Dewan Perwakilan Rakyat DKI Jakarta.

Prediksi hasil pemilihan umum perlu diprediksi dengan akurat, karena hasil prediksi yang akurat sangat penting karena mempunyai dampak pada berbagai macam aspek sosial, ekonomi, keamanan, dan lain-lain (Borisjuk, Borisjuk, Rallings, dan Thrasher, 2005). Bagi para pelaku ekonomi, peristiwa politik seperti pemilu tidak dapat dipandang sebelah mata, mengingat hal tersebut dapat mengakibatkan risiko positif maupun negatif terhadap kelangsungan usaha yang dijalankan.

Metode prediksi hasil pemilihan umum sudah pernah dilakukan oleh peneliti (Rigdon, Jacobson, Sewell, dan Rigdon, 2009) melakukan prediksi hasil pemilihan umum dengan menggunakan metode *Estimator Bayesian*. (Moscato, Mathieson, Mendes, dan Berreta, 2005) melakukan penelitian untuk memprediksi pemilihan presiden Amerika Serikat menggunakan *decision tree*. (Choi dan Han, 1999)

memprediksi hasil pemilihan presiden di Korea dengan metode Decision Tree. (Nagadevara dan Vishnuprasad, 2005) memprediksi hasil pemilihan umum dengan model *classification tree* dan *neural network*. (Borisjuk, Borisjuk, Rallings, dan Thrasher, 2005) yang memprediksi hasil pemilihan umum dengan menggunakan metode *neural network*. Tetapi belum ada yang melakukan perbandingan metode-metode tersebut sehingga belum diketahui metode yang paling akurat.

Untuk itu dalam penelitian ini akan dilakukan perbandingan metode *neural network* dengan metode yang layaknya neuron manusia untuk mencari jalur terbaik dan metode algoritma C4.5 dengan merubah data menjadi pohon keputusan menggunakan rumus perhitungan entropi. Pada penelitian ini, penulis menggunakan aplikasi rapidminer sehingga diperoleh metode dengan akurasi prediksi hasil pemilihan umum Legislatif di DKI Jakarta.

BAHAN DAN METODE

Pemilihan umum (Pemilu) adalah salah satu pilar utama dari sebuah demokrasi, kalau tidak dapat yang disebut yang terutama. Sentralitas dari posisi pemilihan umum dalam membedakan sistem politik yang demokratis dan bukan tampak jelas dari beberapa definisi yang diajukan oleh beberapa peneliti. Salah satu konsepsi modern awal mengenai demokrasi yang diajukan oleh Joseph Schumpeter dan kemudian dikenal dengan mazhab Schumpeterian menempatkan penyelenggaraan pemilihan umum yang bebas dan berkala sebagai kriteria utama bagi sebuah sistem politik untuk dapat disebut sebagai sebuah demokrasi (Santoso, 2004). Dalam sebuah negara demokrasi, pemilihan umum merupakan salah satu pilar utama untuk memilih pemimpin yang nantinya akan mewakili rakyat untuk duduk dipemerintahan mulai dari anggota Dewan Perwakilan Rakyat Daerah (DPRD) tingkat II, Dewan Perwakilan Rakyat Daerah (DPRD) Tingkat I, Dewan Perwakilan Rakyat (DPR) RI dan Dewan Pimpinan Daerah (DPD).

Dalam khazanah demokrasi kontemporer, posisi pemilihan umum memperoleh penguatan. Kajian akademis mengenai demokrasi mengenal dua kategori pemaknaan besar, yaitu konsepsi minimalis dan maksimalis. Demokrasi minimalis atau prosedural dikenakan kepada sistem-sistem politik yang melaksanakan perubahan kepemimpinan secara reguler melalui suatu

mekanisme pemilihan yang berlangsung bebas, terbuka dan melibatkan massa pemilih yang universal. Sedangkan konsep maksimalis adalah pelaksanaan pemilihan umum tidaklah cukup bagi suatu sistem politik untuk mendapatkan gelar demokrasi karena konsep ini mensyaratkan penghormatan terhadap hak-hak sipil yang lebih luas (Santoso, 2004). Pemilu di Indonesia terbagi dari dua bagian, yaitu (Sardini, 2011):

1. Pemilu orde baru

Sistem pemilihannya dilakukan secara proporsional tidak murni, yang artinya jumlah penentuan kursi tidak ditentukan oleh jumlah penduduk saja tetapi juga didasarkan pada wilayah administrasi. Pemilu orde baru dimulai pada tahun 1955 sebagai pemilu pertama yang diselenggarakan di negara Indonesia.

2. Pemilu era reformasi

Dikatakan sebagai pemilu reformasi karena dipercepatnya proses pemilu di tahun 1999 sebelum habis masa kepemimpinan di pemilu tahun 1997. Terjadinya pemilu era reformasi ini karena produk pemilu pada tahun 1997 dianggap pemerintah dan lembaga lainnya tidak dapat dipercaya.

Sistem pemilihan Dewan Perwakilan Rakyat (DPR)/Dewan Perwakilan Rakyat Daerah (DPRD) berdasarkan ketentuan dalam Undang-Undang nomor 10 tahun 2008 pasal 5 ayat 1 sistem yang digunakan dalam pemilihan legislatif adalah sistem proporsional dengan daftar terbuka, sistem pemilihan DPD dilaksanakan dengan sistem distrik berwakil banyak Undang-Undang nomor 10 tahun 2008 pasal 5 ayat 2 tentang pemilihan umum. Menurut Undang-Undang No. 10 tahun 2008 tentang pemilihan umum, Peserta pemilihan anggota DPR/D adalah partai politik peserta Pemilu, sedangkan peserta pemilihan anggota DPD adalah perseorangan. Partai politik peserta Pemilu dapat mengajukan calon sebanyak-banyaknya 120 persen dari jumlah kursi yang diperebutkan pada setiap daerah pemilihan demokratis dan terbuka serta dapat mengajukan calon dengan memperhatikan keterwakilan perempuan sekurang-kurangnya 30 %. Partai Politik Peserta Pemilu diharuskan Undang-Undang untuk mengajukan daftar calon dengan nomor urut (untuk mendapatkan Kursi). Karena itu dari segi pencalonan Undang-Undang No.10 Tahun 2008 tentang Pemilihan umum mengadopsi sistem daftar calon tertutup.

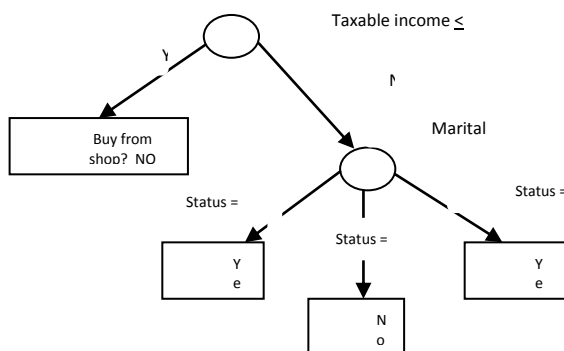
Undang-Undang No.10 Tahun 2008 tentang Pemilihan Umum mengadopsi sistem

proporsional dengan daftar terbuka. sistem proporsional merujuk pada formula pembagian kursi dan/atau penentuan calon terpilih, yaitu setiap partai politik peserta pemilu mendapatkan kursi proporsional dengan jumlah suara sah yang diperolehnya. Penerapan formula proporsional dimulai dengan menghitung Bilangan Pembagi Pemilih (BPP), yaitu jumlah keseluruhan suara sah yang diperoleh seluruh partai politik peserta pemilu pada suatu daerah pemilihan dibagi dengan jumlah kursi yang diperebutkan pada daerah pemilihan tersebut.

Algoritma C4.5

Algoritma C4.5 adalah hasil dari pengembangan algoritma ID3 (*Iterative Dichotomiser*) yang dikembangkan oleh Quinlan (Han dan Kamber 2006). Algoritma ini digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar (Gorunescu, 2011). Sebelumnya diakhir tahun 1970 sampai awal tahun 1980 J. Ross Quinlan, seorang peneliti dibidang *machine learning*, membuat sebuah algoritma *decision tree* yang dikenal dengan ID3 (*Iterative Dichotomiser*).

Algoritma C4.5 atau pohon keputusan mirip sebuah pohon dimana terdapat node internal (bukan daun) yang mendeskripsikan atribut-atribut, setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Pohon keputusan dengan mudah dapat dikonversi ke aturan klasifikasi. Secara umum keputusan pengklasifikasi pohon memiliki akurasi yang baik, namun keberhasilan penggunaan tergantung pada data yang akan diolah.



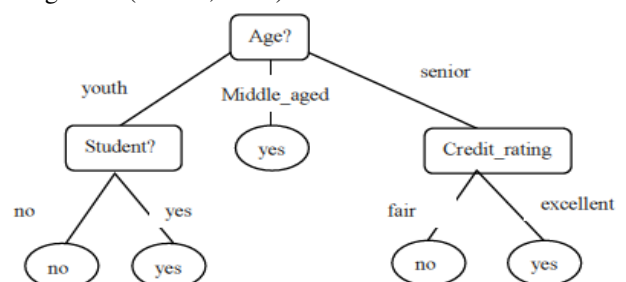
Sumber : Gorunescu, 2011

Gambar 1 Contoh konsep pohon keputusan sederhana

Pada Gambar 1 variabel target untuk pohon keputusan adalah membeli pada toko,

dengan pengklasifikasian ya atau tidak. Variabel *predictor* adalah taxable income (≤ 34.000 atau > 34.000), marital status (married, divorced, atau single). Simpul akar merupakan simpul keputusan, pengujiannya apakah taxable income ≤ 34.000 atau > 34.000 .

Desicion Tree menyerupai struktur *flowchart*, yang masing-masing internal *node*-nya dinyatakan sebagai atribut pengujian, setiap cabang mewakili output dari pengujian, dan setiap *node* daun (*terminal node*) menentukan label *class*. *Node* paling atas dari sebuah pohon adalah *node* akar (Han dan Kamber, 2007). Salah satu metode klasifikasi yang menarik melibatkan konstruksi pohon keputusan, koleksi *node* keputusan, terhubung oleh cabang-cabang, memperpanjang bawah dari simpul akar sampai berakhir di *node* daun. Dimulai di *node root*, yang oleh konvensi ditempatkan di bagian atas dari diagram pohon keputusan, atribut diuji pada *node* keputusan, dengan setiap hasil yang mungkin dihasilkan dalam suatu cabang. Setiap cabang kemudian mengarah baik ke *node* lain keputusan atau ke *node* daun untuk mengakhiri (Larose, 2005).



Sumber : Larose, 2005

Gambar 2. Contoh Decision Tree

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 (kusrini dan Lutfi, 2009), yaitu:

1. Menyiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung gain dari atribut, hitung dahulu nilai entropy yaitu:

$$Entropy(S) = \sum_{i=1}^n -p_i * \log_2 p_i$$

Keterangan:

S : himpunan kasus

A : atribut

N : jumlah partisi

Pi : proporsi dari Si terhadap S

3. Kemudian hitung nilai gain dengan metode *informasi gain*:

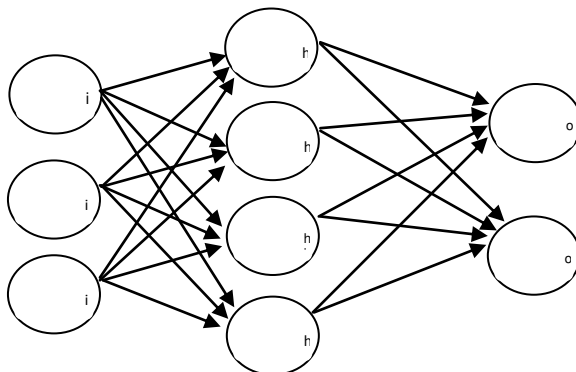
$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

4. Ulangi langkah ke-2 hingga semua tupel terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat:
 - a. Semua tupel dalam node N mendapat kelas yang sama.
 - b. Tidak ada atribut di dalam tupel yang dipartisi lagi.
 - c. Tidak ada tupel di dalam cabang yang kosong.

Neural Network

Neural network adalah suatu sistem pemroses informasi yang memiliki karakteristik menyerupai dengan jaringan saraf biologi pada manusia. *Neural network* didefinisikan sebagai sistem komputasi di mana arsitektur dan operasi diilhami dari pengetahuan tentang sel saraf biologis di dalam otak, yang merupakan salah satu representasi buatan dari otak manusia yang selalu mencoba menstimulasi proses pembelajaran pada otak manusia tersebut (Astuti, 2009).

Neural network dibuat berdasarkan model saraf manusia tetapi dengan bagian-bagian yang lebih sederhana. Komponen terkecil dari *neural network* adalah unit atau yang biasa disebut dengan *neuron* dimana *neuron* tersebut akan mentransformasikan informasi yang diterima menuju *neuron* lainnya (Shukla, Tiwari, dan Kala, 2010).



Sumber : Shukla, Tiwari, dan Kala, 2010
Gambar 3 Model *neural network*

Neural network terdiri dari dua atau lebih lapisan, meskipun sebagian besar jaringan terdiri dari tiga lapisan: lapisan input, lapisan tersembunyi, dan lapisan output (Larose, 2005). Pendekatan *neural network* dimotivasi oleh jaringan saraf biologis. Secara kasar, *neural network* adalah satu set terhubung *input/output* unit, di mana masing-masing sambungan memiliki berat yang terkait dengannya. *Neural network* memiliki beberapa properti yang membuat mereka populer untuk *clustering*. Pertama, *neural network* adalah arsitektur pengolahan *inherent* paralel dan terdistribusi. Kedua, *neural network* belajar dengan menyesuaikan bobot interkoneksi dengan data. Hal ini memungkinkan *neural network* untuk "menormalkan" pola dan bertindak sebagai fitur (atribut) *extractors* untuk kelompok yang berbeda. Ketiga, *neural network* memproses vektor numerik dan membutuhkan pola objek untuk diwakili oleh fitur kuantitatif saja (Gorunescu, 2011).

Neural network terdiri dari kumpulan node (*neuron*) dan relasi. Ada tiga tipe node (*neuron*) yaitu, *input*, *hidden* dan *output*. Setiap relasi menghubungkan dua buah node dengan bobot tertentu dan juga terdapat arah yang menunjukkan aliran data dalam proses (Kusrini dan Luthfi, 2009). Kemampuan otak manusia seperti mengingat, menghitung, mengeneralisasi, adaptasi, diharapkan *neural network* dapat meniru kemampuan otak manusia. *Neural network* berusaha meniru struktur/arsitektur dan cara kerja otak manusia sehingga diharapkan bisa dan mampu menggantikan beberapa pekerjaan manusia. *Neural network* berguna untuk memecahkan persoalan yang berkaitan dengan pengenalan pola, klasifikasi, prediksi dan data mining (Shukla, Tiwari, dan Kala, 2010).

Input node terdapat pada *layer* pertama dalam *neural network*. Secara umum setiap input node merepresentasikan sebuah input parameter seperti umur, jenis kelamin, atau pendapatan. *Hidden node* merupakan node yang terdapat di bagian tengah. *Hidden node* ini menerima masukan dari input node pada *layer* pertama atau dari *hidden node* dari *layer* sebelumnya. *Hidden node* mengombinasikan semua masukan berdasarkan bobot dari relasi yang terhubung, mengkalkulasikan, dan memberikan keluaran untuk *layer* berikutnya. *Output node* mempresentasikan atribut yang diprediksi (Kusrini dan Luthfi, 2009).

Setiap node (*neuron*) dalam *neural network* merupakan sebuah unit pemrosesan.

Tiap node memiliki beberapa masukan dan sebuah keluaran. Setiap node mengkombinasikan beberapa nilai masukan, melakukan kalkulasi, dan membangkitkan nilai keluaran (aktifasi). Dalam setiap node terdapat dua fungsi, yaitu fungsi untuk mengkombinasikan masukan dan fungsi aktifasi untuk menghitung keluaran. Terdapat beberapa metode untuk mengkombinasikan masukan antara lain *weighted sum*, *mean*, *max*, logika OR, atau logika AND (Kusrini dan Luthfi, 2009). Serta beberapa fungsi aktifasi yang dapat digunakan yaitu *heaviside (threshold)*, *step activation*, *piecewise*, *linear*, *gaussian*, *sigmoid*, *hyperbolic tangent* (Gorunescu, 2011).

Salah satu keuntungan menggunakan *neural network* adalah bahwa *neural network* cukup kuat sehubungan dengan data. Karena *neural network* berisi banyak node (*neuron* buatan) dengan bobot ditugaskan untuk setiap koneksi (Larose, 2005).

Algoritma *neural network* mempunyai karakteristik-karakteristik lainnya antara lain (Astuti, 2009),

1. Masukan dapat berupa nilai diskrit atau real yang memiliki banyak dimensi..
2. Keluaran berupa *vektor* yang terdiri dari beberapa nilai diskrit atau real
3. Dapat mengetahui permasalahan secara *black box*, dengan hanya mengetahui nilai masukan serta keluarannya saja.
4. Mampu menangani pembelajaran terhadap data yang memiliki derau(*noise*).
5. Bentuk dari fungsi target pembelajaran tidak diketahui karena hanya berupa bobot-bobot nilai masukan pada setiap *neuron*.
6. Karena harus mengubah banyak nilai bobot pada proses pembelajaran, maka waktu pembelajaran menjadi lama, sehingga tidak cocok untuk masalah-masalah yang memerlukan waktu cepat dalam pembelajaran.
7. *Neural network* hasil pembelajaran tiruan dapat dijalankan dengan tepat.

Penelitian di bidang *neural network* dimulai pada masa komputer digital. McCulloch dan Pitts (1943) mengemukakan model matematika pertama untuk *neural network*. Rosenblatt (1962) mengemukakan model *perceptron* dan algoritma pembelajaran pada tahun 1962 (Alpaydin, 2010). Minsky dan Papert (1969) menunjukkan keterbatasan *single layer perceptron* untuk menyelesaikan masalah yang *nonlinearly separable*. Kemudian Rumelhart, Hinton, and Williams (1986)

yang mempresentasikan algoritma *backpropagation* untuk *multilayer perceptron* yang dapat menyelesaikan masalah yang *nonlinearly separable* (Han dan Kamber, 2007). Aplikasi *neural network* telah banyak dimanfaatkan untuk berbagai kepentingan seperti di bidang Elektronik, Otomotif, Perbankan, Sistem penerbangan udara, Dunia hiburan, transportasi publik, telekomunikasi, bidang Kesehatan, Keamanan, bidang Robotika, Asuransi, Pabrik, Financial, Suara, Pertambangan dan sistem pertahanan (Astuti, 2009).

Algoritma yang paling populer pada algoritma *neural network* adalah algoritma *backpropagation*. Algoritma pelatihan *backpropagation* atau ada yang menterjemahkan menjadi propagasi balik pertama kali dirumuskan oleh Paul Werbos pada tahun 1974 dan dipopulerkan oleh Rumelhart bersama McClelland untuk dipakai pada *neural network*. Metode *backpropagation* pada awalnya dirancang untuk *neural network feedforward*, tetapi pada perkembangannya, metode ini diadaptasi untuk pembelajaran pada model *neural network* lainnya (Astuti, 2009). Salah satu metode pelatihan terawasi pada *neural network* adalah metode *backpropagation*, di mana ciri dari metode ini adalah meminimalkan *error* pada output yang dihasilkan oleh jaringan.

Metode algoritma *backpropagation* ini banyak diaplikasikan secara luas. *backpropagation* telah berhasil diaplikasikan di berbagai bidang, antaranya bidang finansial, pengenalan pola tulisan tangan, pengenalan pola suara, sistem kendali, pengolah citra medika. *backpropagation* berhasil menjadi salah satu metode komputasi yang handal. Algoritma *backpropagation* mempunyai pengaturan hubungan yang sangat sederhana yaitu: jika keluaran memberikan hasil yang salah, maka penimbang (*weight*) dikoreksi supaya galatnya dapat diperkecil dan respon jaringan selanjutnya diharapkan akan mendekati nilai yang benar. Algoritma ini juga berkemampuan untuk memperbaiki penimbang pada lapisan tersembunyi (*hidden layer*) (Purnomo dan Kurniawan, 2006).

Inisialisasi awal bobot jaringan *backpropagation* yang terdiri atas lapisan input, lapisan tersembunyi, dan lapisan output (Astuti, 2009). Tahap pelatihan *backpropagation* merupakan langkah untuk melatih suatu *neural network* yaitu dengan cara melakukan perubahan penimbang (sambungan antar lapis yang membentuk *neural network* melalui masing-masing

unitnya). Sedangkan penyelesaian masalah akan dilakukan jika proses pelatihan tersebut telah selesai, fase ini disebut dengan fase *mapping* atau proses pengujian/testing.

Langkah pembelajaran dalam algoritma *bakpropagation* adalah sebagai berikut (Myatt, 2007):

1. Inisialisasi bobot jaringan secara acak (biasanya antara -0.1 sampai 1.0)
2. Untuk setiap data pada data *training*, hitung input untuk simpul berdasarkan nilai input dan bobot jaringan saat itu, menggunakan rumus:

$$\text{Input } j = \sum_{i=1}^n O_i W_{ij} + \theta_j$$

3. Berdasarkan input dari langkah dua, selanjutnya membangkitkan output.

untuk simpul menggunakan fungsi aktivasi sigmoid:

$$\text{Output} = \frac{1}{1 + e^{-\text{input}}}$$

4. Hitung nilai *Error* antara nilai yang diprediksi dengan nilai yang sesungguhnya menggunakan rumus:

$$\text{Error}_j = \text{output}_j \cdot (1 - \text{output}_j) \cdot (\text{Target}_j - \text{Output}_j)$$

5. Setelah nilai *Error* dihitung, selanjutnya dibalik ke *layer* sebelumnya (*bakpropagation*). Untuk menghitung nilai *Error* pada *hidden layer*, menggunakan rumus:

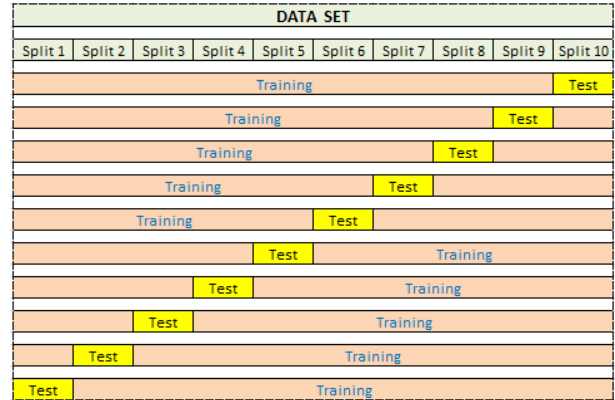
$$\text{Error}_j = \text{Output}_j (1 - \text{Output}_j) \sum_{k=1}^n \text{Error}_k W_{jk}$$

6. Nilai *Error* yang dihasilkan dari langkah sebelumnya digunakan untuk memperbarui bobot relasi menggunakan rumus:

$$W_{ij} = W_{ij} + l \cdot \text{Error}_j \cdot \text{Output}_i$$

Pengujian K-Fold Cross Validation

Cross Validation adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi (Han dan Kamber, 2007). Dengan menggunakan *cross validation* akan dilakukan percobaan sebanyak k. Data yang digunakan dalam percobaan ini adalah data *training* untuk mencari nilai *error rate* secara keseluruhan. Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi. Dalam penelitian ini nilai k yang digunakan berjumlah 10 atau 10-fold *Cross Validation*.



Sumber : Han dan Kamber, 2007

Gambar 4. Ilustrasi 10 Fold Cross Validation

Pada gambar 4 terlihat bahwa tiap percobaan akan menggunakan satu data *testing* dan k-1 bagian akan menjadi data *training*, kemudian data *testing* itu akan ditukar dengan satu buah data *training* sehingga untuk tiap percobaan akan didapatkan data *testing* yang berbeda-beda.

Confusion Matrix

Confusion matrix memberikan keputusan yang diperoleh dalam *training* dan *testing*, *confusion matrix* memberikan penilaian *performance* klasifikasi berdasarkan objek dengan benar atau salah (Gorunescu, 2011).

Confusion matrix berisi informasi aktual (*actual*) dan prediksi (*predicted*) pada sistem klasifikasi. Berikut tabel penjelasan tentang confusion matrix.

Tabel 1. Confusion Matrix

Sumber : Gorunescu, 2011

Classifica tion	Predicted Class		
		Class = Yes	Class = No
Observed Class	Class = Yes	A(True Positif-tp)	B(False negatif- fn)
	Class = No	C(False positif- fp)	D(true negative- tn)

Keterangan:

True Positive (tp) = proporsi positif dalam data set yang diklasifikasikan positif

True Negative (tn) = proporsi negative dalam data set yang diklasifikasikan negative

False Positive (fp) = proporsi negatif dalam data set yang diklasifikasikan positif

$FalseNegative(fn)$ = proporsi negative dalam data set yang diklasifikasikan negatif

Berikut adalah persamaan model *confusion matrix*:

- a. Nilai akurasi (acc) adalah proporsi jumlah prediksi yang benar.

Dapat dihitung dengan menggunakan persamaan:

$$acc = \frac{tp + tn}{tp + tn + fp + fn}$$

- b. Sensitivity digunakan untuk membandingkan proporsi tp terhadap tupel yang positif, yang dihitung dengan menggunakan persamaan:

$$Sensitivity = \frac{tp}{tp + fn}$$

- c. Specificity digunakan untuk membandingkan proporsi tn terhadap tupel yang negatif, yang dihitung dengan menggunakan persamaan:

$$Specificity = \frac{tn}{tn + fp}$$

- d. PPV (*positive predictive value*) adalah proporsi kasus dengan hasil diagnosa positif, yang dihitung dengan menggunakan persamaan:

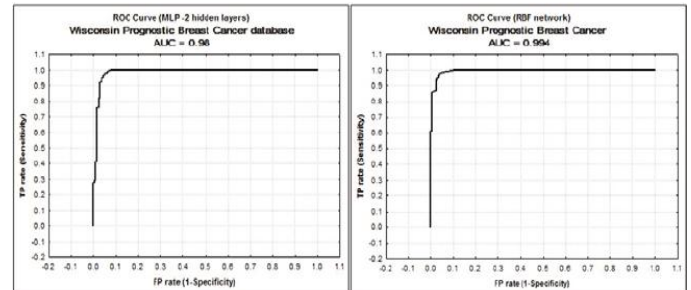
$$PPV = \frac{tp}{tp + fp}$$

- e. NPV (*negative predictive value*) adalah proporsi kasus dengan hasil diagnosa negatif, yang dihitung dengan menggunakan persamaan:

$$NPV = \frac{tn}{tn + fn}$$

Curve ROC

Curve ROC (*Receiver Operating Characteristic*) adalah cara lain untuk mengevaluasi akurasi dari klasifikasi secara visual (Vercellis, 2009). Sebuah grafik ROC adalah plot dua dimensi dengan proporsi positif salah (fp) pada sumbu X dan proporsi positif benar (tp) pada sumbu Y. Titik (0,1) merupakan klasifikasi yang sempurna terhadap semua kasus positif dan kasus negatif. Nilai positif salah adalah tidak ada (fp = 0) dan nilai positif benar adalah tinggi (tp = 1). Titik (0,0) adalah klasifikasi yang memprediksi setiap kasus menjadi negatif {-1}, dan titik (1,1) adalah klasifikasi yang memprediksi setiap kasus menjadi positif {1}. Grafik ROC menggambarkan *trade-off* antara manfaat (*true positives*) dan biaya (*false positives*). Berikut tampilan dua jenis kurva ROC (*discrete* dan *continous*).



Sumber : Gorunescu, 2011

Gambar 5. Grafik ROC (*discrete* dan *continous*)

Pada Gambar 2.4 garis diagonal membagi ruang ROC, yaitu:

1. (a) poin diatas garis diagonal merupakan hasil klasifikasi yang baik.
2. (b) point dibawah garis diagonal merupakan hasil klasifikasi yang buruk.

Dapat disimpulkan bahwa, satu point pada kurva ROC adalah lebih baik dari pada yang lainnya jika arah garis melintang dari kiri bawah ke kanan atas didalam grafik. Tingkat akurasi dapat di diagnosa sebagai berikut (Gorunescu, 2011):

Akurasi 0.90 – 1.00 = *Excellent classification*

Akurasi 0.80 – 0.90 = *Good classification*

Akurasi 0.70 – 0.80 = *Fair classification*

Akurasi 0.60 – 0.70 = *Poor classification*

Akurasi 0.50 – 0.60 = *Failure*

METODE PENELITIAN

Menurut Sharp et al (Dawson, 2009) penelitian adalah mencari melalui proses yang metodis untuk menambahkan pengetahuan itu sendiri dan dengan yang lainnya, oleh penemuan fakta dan wawasan tidak biasa. Pengertian lainnya, penelitian adalah sebuah kegiatan yang bertujuan untuk membuat kontribusi orisinal terhadap ilmu pengetahuan (Dawson, 2009).

Penelitian ini adalah penelitian eksperimen dengan metode penelitian sebagai berikut

1. Pengumpulan data

Pada pengumpulan data dijelaskan tentang bagaimana dan darimana data dalam penelitian ini didapatkan, ada dua tipe dalam pengumpulan data, yaitu pengumpulan data primer dan pengumpulan data sekunder. Data primer adalah data yang dikumpulkan pertama kali untuk melihat apa yang sesungguhnya terjadi

Data sekunder adalah data yang sebelumnya pernah dibuat oleh seseorang baik di terbitkan atau tidak (Kothari, 2004). Dalam pengumpulan data primer dalam penelitian ini menggunakan metode observasi dan interview, dengan menggunakan data-data yang berhubungan dengan pemilu ditahun 2009. Data yang didapat dari KPUD Jakarta adalah data pemilu tahun 2009 dengan jumlah data sebanyak 2268 *record*, terdiri dari 11 variabel atau atribut. Adapun variabel yang digunakan yaitu no urut partai, nama partai, suara sah partai, no urut caleg, nama caleg, jenis kelamin, kota administrasi, daerah pemilihan, suara sah caleg, jumlah perolehan kursi. Sedangkan variabel tujuannya yaitu hasil pemilu.

2. Pengolahan awal data

Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 2.268 data, namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data (*preparation data*). Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan sebagai berikut (Vercellis, 2009):

- a. Data validation, untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*).
- b. Data integration and transformation, untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan kedalam *software Rapidminer*. Tabel kategorikal atribut terlihat pada table 3.2.
- c. Data size reduction and discretization, untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informative.

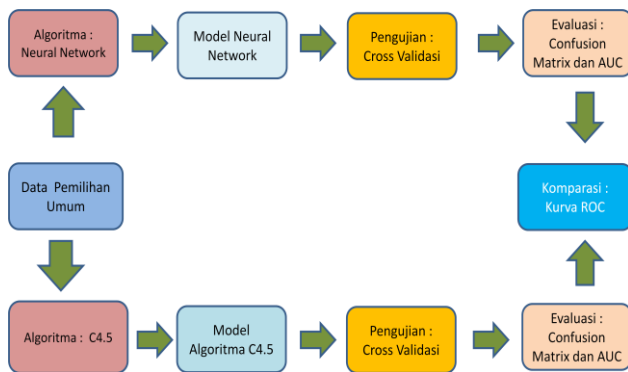
Tabel 2. Atribut yang digunakan

No	Atribut	Nilai
1	Nama Partai	Nama Partai Peserta Pemilu Legislatif
2	Nama Calon Legislatif	Nama Peserta Pemilu Legislatif dari Partai Politik
3	Jenis kelamin	Jenis kelamin caleg (L/ P)
4	Kota Administrasi	Nama 5 Kota administrasi di DKI Jakarta
5	No.urut parpol	Nomor urut partai politik dari KPU
6	Suara sah partai	Jumlah akumulasi perolehan suara partai politik
7	Juml.perolehan kursi	Jumlah perolehan kursi partai
8	Daerah pemilihan	Daerah pemilihan (1/2/3/4/5)
9	No.urut caleg	Nomor urut caleg di partai politik tersebut
10	Suara sah caleg	Jumlah suara sah caleg

Sumber : Penelitian Tahun 2014

3. Model yang diusulkan

Dalam penelitian ini akan dilakukan analisis komparasi menggunakan tiga metode klasifikasi data mining. Metode yang diusulkan untuk pengolahan data mahasiswa adalah penggunaan algoritma C4.5, *naïve bayes* dan *neural network*. Data diolah sesuai dengan algoritmanya masing-masing, yakni data mahasiswa diolah menggunakan metode algoritma C4.5, *naïve bayes* dan *neural network*, setelah diolah dan menghasilkan model, maka terhadap model yang dihasilkan tersebut dilakukan pengujian menggunakan *k-fold cross validation*, kemudian dilakukan evaluasi dan validasi hasil dengan *confusion matrix* dan kurva ROC. Tahap selanjutnya adalah membandingkan hasil akurasi dan AUC dari setiap model, sehingga diperoleh model dari metode klasifikasi yang mana yang memperoleh nilai akurasi dan AUC tertinggi. Dalam tahapan ini akan dilakukan beberapa langkah-langkah metode yang diusulkan data yaitu seperti berikut:



Sumber : Dawson, 2009

Gambar 6. Model yang diusulkan

Hasil pengujian dengan akurasi yang paling tinggi adalah metode yang akan digunakan untuk prediksi kelulusan mahasiswa ini. Berikut gambaran karakteristik dari masing-masing metode:

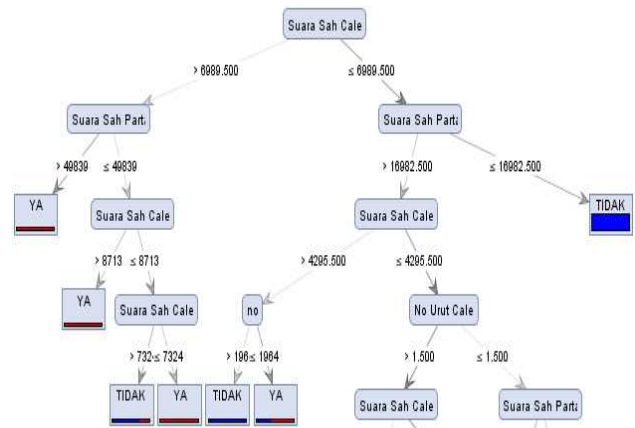
- Algoritma C4.5 yaitu salah satu algoritma dalam metode *decision tree* yang merubah data menjadi pohon keputusan menggunakan rumus perhitungan entropi.
- Neural network* yaitu metode layaknya *neuron* manusia untuk mencari jalur terbaik.

HASIL DAN PEMBAHASAN

Decision Tree

Algoritma C4.5 dan pohon keputusan (*decision tree*) merupakan dua model yang tidak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhkan algoritma C4.5. Tujuan utama dari menganalisis data dengan menggunakan algoritma *Decision Tree* ini adalah ingin mendapatkan rule (Witten, Frank, dan Hall, 2011) yang akan dimanfaatkan untuk pengambilan keputusan pada data baru.

Berikut pohon keputusan dan algoritma C4.5 yang terbentuk dari pengolahan data pemilu legislatif DKI Jakarta.



Sumber : Penelitian tahun 2014

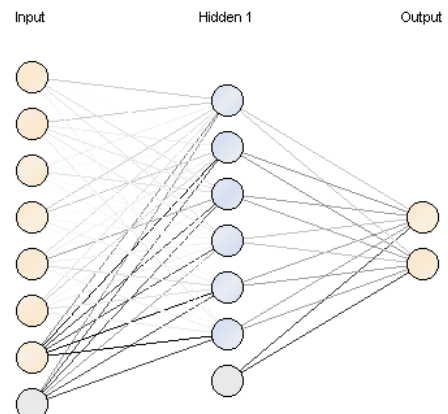
Gambar 7. Pohon keputusan algoritma C4.5

Berdasarkan eksperimen yang penulis lakukan, hasil terbaik menunjukkan bahwa, tingkat akurasi dengan menggunakan algoritma C4.5 adalah sebesar 97,84% dan AUC dari algoritma C4.5 sebesar 0.970.

Metode Neural Network

Algoritma *neural network* adalah algoritma untuk pelatihan *supervised* dan didesain untuk operasi pada *feed forward* multilapis. Algoritma *neural network* bisa dideksripsikan sebagai berikut: ketika jaringan diberikan pola masukan sebagai pola pelatihan maka pola tersebut menuju ke unit-unit pada lapisan tersembunyi untuk diteruskan ke unit-unit lapisan terluar. Hasil terbaik pada *eksperiment* adalah dengan *accuracy* yang dihasilkan sebesar 98.50 dan AUCnya 0.982.

Dari eksperimen terbaik di atas maka didapat arsitektur *neural network* dengan menghasilkan enam *hidden layer* dengan tujuh atribut *input layer* dan dua *output layer*. Gambar arsitektur *neural network* terlihat pada gambar 2 seperti di bawah ini



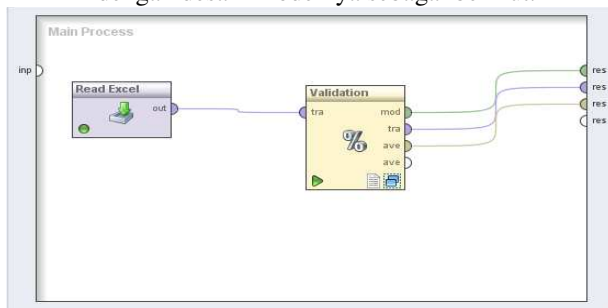
Sumber : Penelitian tahun 2014

Gambar 8. Arsitektur *neural network*

Analisa Evaluasi dan Validasi Model

Hasil dari pengujian model yang dilakukan adalah memprediksi hasil pemilu legislatif DKI Jakarta 2009 adalah dengan membandingkan *algoritma C4.5* dan *neural network* untuk menentukan nilai *accuracy* dan *AUC*.

Dalam menentukan nilai tingkat keakurasian dalam *algoritma C4.5* dan *algoritma neural network*, Metode pengujiannya menggunakan *cross validation* dengan desain modelnya sebagai berikut.



Sumber : Penelitian tahun 2014

Gambar 9. Pengujian *cross validation*

Dari hasil pengujian diatas, baik evaluasi menggunakan *counfusion matrix* maupun *ROC curve* terbukti bahwa hasil pengujian *algoritma neural network* berbasis PSO memiliki nilai akurasi yang lebih tinggi dibandingkan dengan *algoritma neural network*.

Dari hasil pengujian *algoritma decision tree* dan *neural network*, baik evaluasi menggunakan *counfusion matrix* maupun *ROC curve* terbukti bahwa hasil pengujian *algoritma neural network* memiliki nilai akurasi yang lebih tinggi dibandingkan dengan *algoritma decision tree*.

Nilai akurasi untuk model *algoritma neural network* sebesar 98.50 % dan nilai *AUC* sebesar 0.982. sedangkan akurasi untuk metode *decision tree* sebesar 97.84 % dengan *AUC* sebesar 0.970, dapat dilihat pada Tabel 1 dibawah ini:

Tabel 3. Perbandingan nilai akurasi

	Accuracy	AUC
Decision Tree	97.84	0.970
Neural Network	98.50	0.982

Sumber : Penelitian tahun 2014

KESIMPULAN

Dari Dari hasil perbandingan *algoritma decision tree* dan *algoritma neural network* diperoleh hasil yaitu nilai akurasi yang dapat untuk *algoritma decision tree* sebesar 97.84 %. Sedangkan untuk *algoritma neural network* nilai akurasi yang dihasilkan dari *algoritma ini* sebesar 98.50 %. Dari hasil tersebut didapatkan selisih untuk nilai akurasi antara *algoritma decision tree* dan *algoritma neural network* yaitu 0.66 %. Sementara untuk evaluasi menggunakan *ROC curve* untuk kedua model yaitu, untuk model *algoritma neural network* nilai *AUC* adalah 0.982 dengan tingkat diagnosa *Excellent Classification*, dan untuk model *algoritma decision tree* nilai *AUC* adalah 0.970 dengan tingkat diagnosa *Excellent Classification*. Dari evaluasi *ROC curve* tersebut terlihat bahwa model *neural network* lebih tinggi jika dibandingkan dengan *algoritma decision tree*. Dari hasil tersebut didapatkan selisih antara kedua model yaitu 0.012. Sehingga dapat disimpulkan bahwa *algoritma neural network* lebih akurat dalam memprediksi hasil pemilu legislatif jika dibandingkan dengan *algoritma decision tree*.

DAFTAR PUSTAKA

- Astuti, E. D. (2009). Pengantar Jaringan Saraf Tiruan. Wonosobo: Star Publishing.
- Borisyuk, R., Borisyuk, G., Rallings, C., dan Thrasher, M. (2005). *Forecasting the 2005 General Election: A Neural Network Approach. The British Journal of Politics dan International Relations Volume 7, Issue 2*, 145-299.
- Choi, J. H., dan Han, S. T. (1999). *Prediction of Elections Result using Discrimination of Non-Respondents: The Case of the 1997 Korea Presidential Election*.
- Dawson, C. W. (2009). *Projects in Computing and Information System A Student's Guide*. England: Addison-Wesley.
- Gorunescu, F. (2011). *Data Mining Concepts, Model and Technique*. Berlin: Springer.
- Han, J., dan Kamber, M. (2007). *Data Mining Concepts and Technique*. Morgan Kaufmann publisher

- Kothari, C. R. (2004). *Research Methodology methodes and Technique*. India: New Age Interntional.
- Kusrini, dan Luthfi, E. T. (2009). *Algoritma Data mining*. Yogyakarta: Andi.
- Larose, D. T. (2005). *Discovering Knowledge in Data*. Canada: Wiley Interscience.
- Moscato, P., Mathieson, L., Mendes, A., dan Berreta, R. (2005). *The Electronic Primaries: Prediction The U.S. Presidential Using Feature Selection with safe data. ACSC '05 Proceeding of the twenty-eighth Australian conference on Computer Science Volume 38* , 371-379.
- Myatt, G. J. (2007). *Making Sense of Data A Practical Guide to Exploratory Data Analysis and Data Mining*. New Jersey: A John Wiley dan Sons, inc., publication.
- Nagadevara, dan Vishnuprasad. (2005). *Building Predictive models for election result in india an application of classification trees and neural network. Journal of Academy of Business and Economics Volume 5* .
- Purnomo, M. H., dan Kurniawan, A. (2006). *Supervised Neural Network*. Suarabaya: Garaha Ilmu.
- Rigdon, S. E., Jacobson, S. H., Sewell, E. C., dan Rigdon, C. J. (2009). *A Bayesian Prediction Model For the United State Presidential Election. American Politics Research volume.37* , 700-724.
- Santoso, T. (2004). Pelanggaran pemilu 2004 dan penanganannya. *Jurnal demokrasi dan Ham* , 9-29.
- Sardini, N. H. (2011). *Restorasi penyelenggaraan pemilu di Indonesia*. Yogyakarta: Fajar Media Press.
- Shukla, A., Tiwari, R., dan Kala, R. (2010). *Real Life Application of Soft Computing*. CRC Press.
- Undang-Undang RI No.10. (2008).
- Vercellis, C. (2009). *Business Intelligence : Data Mining and Optimization for Decision Making*. John Wiley dan Sons, Ltd.
- Witten, H. I., Eibe, F., dan Hall, A. M. (2011). *Data Mining Machine Learning Tools and Techiques*. Burlington: Morgan Kaufmann Publisher.

BIODATA PENULIS

Mohammad Badrul, M.Kom. Penulis adalah Dosen Tetap di STMIK Nusa Mandiri Jakarta. Penulis Kelahiran di Bangkalan 01 Januari 1984. Penulis menyelesaikan Program Studi Strata 1 (S1) di Kampus STMIK Nusa Mandiri Prodi Sistem Informasi dengan gelar S.Kom pada tahun 2009 dan menyelesaikan progarm Srata 2 (S2) di Kampus yang sama dengan Prodi ilmu Komputer dengan gelar M.Kom pada tahun 2012. Selain mengajar, Penulis juga aktif dalam membimbing mahasiswa yang sedang melakukan penelitian khususnya di tingkat Strata 1 dan penulis juga terlibat dalam tim konsorsium di Jurusan Teknik Informatika STMIK Nusa Mandiri Jakarta untuk penyusunan bahan ajar. Saat ini penulis memiliki Jabatan Fungsional Asisten ahli di kampus STMIK Nusa Mandiri Jakarta. Penulis tertarik dalam bidang kelimuan Data mining, Jaringan komputer, *Operating sistem* khususnya *open source, Database, Software engineering* dan *Research Metode*.

